Generating MEDLINE Search Strategies Using a Librarian Knowledge-Based System

Ping Peng, Ph.D; Anthony Aguirre, M.S., M.L.S.;
Stephen B. Johnson, Ph.D.; James J. Cimino, M.D.

Center for Medical Informatics
Columbia-Presbyterian Medical Center
New York, New York 10032

We describe a librarian knowledge-based system that generates a search strategy from a query representation based on a user's information need. Together with the natural language parser AQUA, the system functions as a human/computer interface, which translates a user query from free text into a BRS Onsite search formulation, for searching the MEDLINE bibliographic database. In the system, conceptual graphs are used to represent the user's information need. The UMLS Metathesaurus and Semantic Net are used as the key knowledge sources in building the knowledge base.

INTRODUCTION

One of the main tasks involved in searching an information database is to express the user's information need in the query language of the retrieval system. Generally, users express their information need in natural language or by using a restricted artificial language which is defined for particular system. Constructing a good query formulation requires an understanding of the language and internal characteristics of the system. As more and more databases become available, the learning required to achieve fluency with more than a few sources becomes impractical.

The quality of the query formulation generally can be improved with the assistance of a librarian, whose knowledge of the database, retrieval language syntax, and indexing terms plays a key role. Initial investigations have used librarian knowledge to respond to constrained user queries [1]. Our system attempts to encapsulate the librarian's expertise and to use that knowledge, together with the National Library of Medicine's (NLM's) Unified Medical Language System (UMLS) Metathesaurus [2,3,4] and Semantic Network [5], to build strategies for searching Medline from users' natural language queries.

Overall, the system serves as an interface to a MEDLINE bibliographic database. It consists of the natural language front end, the MEDLINE search strategy generator and a communications program to allow the strategy to transmit the strategy to the MEDLINE search engine. The first step in the process uses a natural language parser called AQUA (A QUery Analyzer) [6] to identify concepts and relations in the query and represent them in a standard notation (conceptual graphs) which is independent of the syntax of any one information source. This paper describes the next step: using a knowledge base of librarian expertise to translate the intermediate notation into MEDLINE strategies.

SEARCH STRATEGY GENERATOR INPUT

The key components in a user's information request are *concepts*, appearing as terms of one or more words, for example, "tuberculosis" or "Hansen's disease". Similar terms can be grouped into recognizable classes. For example, "tuberculosis" and "Hansen's disease" are both in the class "disease" (or, in the UMLS, "Disease or Syndrome"). By grouping terms into a manageable number of classes, we are able to develop efficient knowledge operations that are class-specific. We represent user concepts as a semantic type and referent pair. The referent can be one or more terms in the semantic type. For example, the concept "tuberculosis" is: **['disease or syndrome': tuberculosis]**; two concepts of the same type encountered in the query are included in the same referent. For example, "tuberculosis and Hansen's disease": **['disease or syndrome': {tuberculosis, Hansen's disease}]**.

Concepts in a user's query are usually related grammatically. We represent a relation/concept structure as a *semantic triple*: [concept1, relation, concept2]; where concepts 1 and 2 are instantiated semantic types. A typical triple might be:

[['bacterium': 'mycobacterium tuberculosis'], causes,
['disease or syndrome': 'tuberculosis']].

The semantic types and relations are drawn from the UMLS Semantic Net. The referents to each semantic type are instantiated with words from the user's query. Wherever possible, the referents are translated to their corresponding MeSH terms, found in the UMLS Metathesaurus. If the user term is the same as the MeSH term, the MeSH term is used (indicated here with a "#"); if the user term is literally different from the MeSH term, both the user term and its corresponding MeSH term are used as the referents (for example, "Hansen's disease" would be represented by both that string and the corresponding MeSH term "leprosy").

In general, the information need expressed by a user's query is described by one or more semantic triples. These semantic triples are related in either parallel or nested structure, necessitating a sophisticated knowledge representation formalism. AQUA[6] uses a semantic grammar to parse user queries to extract the concepts expressed and to organize these concepts and their relations into conceptual graphs [7]. Conceptual graphs derive from a system of logic based on the existential graph [8,9] of Charles Sanders Peirce and the semantic networks of artificial intelligence. Conceptual graphs provide a formalism that is logically precise, humanly readable, and computationally tractable; they have been used successfully to represent the literal meaning of sentences in natural language processing [7]. The conceptual graph of the above semantic triple is:

['bacterium': {#'mycobacterium tuberculosis'}]
(causes -> ['disease or syndrome': {#tuberculosis}]).

Graphs such as these serve as the formal representation of user information needs which are processed by the search strategy generator.

THE LIBRARIAN KNOWLEDGE BASE

We classify librarian expertise into four different types: interactions with the user to determine precisely what information is needed; selection of the information source appropriate to the information need; generation of a search strategy which is syntactically correct for the selected information source; and verification that the semantic content of the strategy is appropriate to the information need. The first type of expertise is incorporated into the AQUA parser. The second type is not utilized in the current project, since MEDLINE is used as the sole information. The third (syntactic) and fourth (semantic) types of expertise are encoded in the grammar rules used by the search strategy generator.

Syntactic Expertise

The syntactic component of the generator encodes rules for the conversion of conceptual graphs into BRS search strategies. The MEDLINE BRS query guidelines [10] can be expressed by the Definite Clause Grammar (DCG) formalism (a generalization of context-free grammars that are executable) in programming language PROLOG. Usually, a DCG is used for parsing a sentence string into a structure such as a conceptual graph. However, using the unification feature of logic programming[11] the computation can be reversed, translating a conceptual graph into a string of words in the syntax of a given language. A fragment of the syntax for converting conceptual graphs to BRS search strategies is:

```
search -> orclauses.
orclauses -> clause.
orclauses -> clause, [or], orclauses.
clause -> entity, [Op], entity.
entity -> term.
entity -> ['('], orterms, [')'].
orterms -> term.
orterms -> term, [or], orterms.
```

The first rule states that a "search" may consists of a sequence of clauses; the second rule states that the sequence of clauses can be a single clause alone or a single clause followed by an "or" operator followed by another sequence of clauses. Thus, a search is defined recursively as one or more clauses separated by "or". Similarly, a clause is defined as one or more "entities" separated by operators, where an entity can be a single "term" or a list of terms in parentheses and separated by "or". The symbol "Op" refers to the BRS conjunction operators AND, SAME or WITH.

The simple grammar above can define an infinite set of strategies; there is no limit on the number of clauses which can be included in a valid strategy. The actual number of clauses generated is determined by the given conceptual graph. In order to provide accurate placement of concepts and relations in their proper positions in the search strategy, the grammar was extended by adding logic variables. The values of the variables are used for deterministic choices in generating a search strategy. For example, the second and third rules in the above grammar are provided with variables as follows:

```
orclauses(Concept, [Relation]) ->
    clause(Concept,Relation).
orclauses(Concept, [Relation | Relations]) ->
    clause(Concept, Relation), [or],
    orclauses(Concept, Relations).
```

The syntactic grammar provides a mechanism for converting conceptual graphs to BRS search strategies, but it provides no information about how the strategies should be constructed for specific purposes. For example, logic variables are helpful for selecting a rule for generating a search strategy, but are not useful in determining a value of the operator variable, [Op]. The generation of search-specific information, such as operators, requires additional rules derived from librarian expertise.

Semantic Expertise

The generation of appropriate search strategies is something of an art. The use of a particular operator or subheading with one search may retrieve a manageable number of citations, while producing an unwieldy number for a second search and none with a third. We elicited from librarians the specific operators and subheadings they would use when confronted with a search that involved terms from two particular semantic types, related in a particular way. From this information, we were able to develop search patterns that could be incorporated into the query. For example, it was learned from the librarians that when searching for citations about bacterial causes of diseases, the bacterial term and the disease term should be combined using the "and" operator. In addition, the Subheading "et" (for "etiology") should be appended to the disease term to provide a more specific search strategy. This information was encoded as the pattern:

```
pattern([causes,'bacterium','disease or syndrome'],
    and,[' ','-et']).
```

The patterns can be applied by incorporating references to them in the syntactic grammar. For example, the rule which defines search clauses:

```
clause -> entity, [Op], entity.
```

is modified to the form:

```
clause([Type1:Cpt1],[Rel,Dir,[[[Type2:Cpt2]]]]) ->
    {pattern([Rel,Type1,Type2],Op,[Sh1,Sh2])},
    entity(Cpt1,Sh1),[Op],entity(Cpt2,Sh2).
```

Here, "Type1", "Type2", "Cpt1", "Cpt2", "Rel", "Dir", "Op", "Sh1" and "Sh2" are logic variables, which are instantiated with a user query. The pairs [Type1:Cpt1] and [Type2:Cpt2], represent the two concepts referents in the triple; "Rel" represents the relation between the two concepts; "Dir" indicates the direction of the relation; "Op" represents the operator that connects the terms corresponding to the two concepts; "Sh1" and "Sh2" provide subheadings.

Taken together, the syntactic grammar and the semantic patterns are sufficient to generate functional BRS search strategies from conceptual graphs. For example, the conceptual graph

```
['bacterium': {#'mycobacterium tuberculosis'}]
 (causes -> ['disease or syndrome': {#tuberculosis}]).
```

can be converted to a clause by partially instantiating the clause rule as:

```
clause(['bacterium':{#'mycobacterium tuberculosis'}],
 [causes,'->',[[['disease or syndrome':
 {#tuberculosis}]]]]) ->
    {pattern([causes,'bacterium','disease or syndrome']
    ,Op,[Sh1,Sh2])},
    entity({#'mycobacterium tuberculosis'},Sh1),
    [Op],entity({#tuberculosis},Sh2).
```

This partial instantiation is sufficient for the above pattern to be evoked. This evocation, in turn, allows the further binding of "Op" to "and", "Sh1" to the null string and "Sh2" to the MeSH Subheading "et".

Although MEDLINE is used as the sole search source, the scheme of the computational model for generating a query formulation is intended to be independent of a particular information retrieval language and system.

SEARCH STRATEGY GENERATION

The search strategy generator applies the librarian knowledge encoded in the rules to generate text strings from the conceptual graphs produced by AQUA. For example, the following user query was obtained from a set of queries collected by the NLM:

"Impedance plethysmography or rheography of brain and eye. called also rheoencephalography and rheoophthalmography."

AQUA parses this sentence to produce the graph:

```
['diagnostic procedure':{#'plethysmography
 impedance', #rheography,rheoencephalography,
 rheoophthalmography}]
 (associated_with ->
    ['body part,organ, or organ component':
        {#brain,#eye}])]
```

The search strategy generator produces the string:

"(plethysmography-impedance.de. or rheography.de. or rheoencephalography or rheoophthalmography) and (brain.de. or eye.de.)"

(In the search strategy, ".de." indicates restriction of terms to "descriptor" fields of BRS citation records.)

The BRS search engine takes the above search strategy as the input and conducts the search in MEDLINE. As a result, 42 documents are reported to be found in MEDLINE. For this query, most of the retrieved documents are relevant. A search strategy for the same user query, created manually by a librarian, obtains 78 documents from MEDLINE, 35 of which are in common with the automated search.

CURRENT STATUS

The system has been implemented in PROLOG. The Definite Clause Grammar provides a complete description of BRS Onsite usages and generates any legitimate form of a search formulation. The patterns are based on 417 semantic triples extracted from 339 user queries (445 sentences). There are currently 16 DCG grammar rules and 51 patterns; the Prolog program itself comprises 23 predicate functions.

The system runs, together with AQUA, on an IBM RS/6000 workstation running AIX. A scripting language (Expect) is used to allow a user to input natural language queries into AQUA and then transfer the resultant conceptual graphs from AQUA to the search strategy generator. The script then transfers the resulting strategies, via TCP/IP protocol, to a BRS/Onsite search engine running on an IBM 3090/300. At that point, the script allows direct interaction between the user and BRS to review search results and modify search strategies.

DISCUSSION

There have been several efforts to develop methods for automatically formulating search strategies and carrying out the retrieval for searching the medical literature. A number of attempts have been made to facilitate access to information retrieval systems. NLM's Coach expert search system [12] revises failed MEDLINE searches by trying out different combinations of three- and four-term Boolean "AND" searches in order to improve document retrieval. IQW [13] provides an interface to several information sources that allows a user to select a query, chooses an appropriate database and formulates an initial search strategy. CHARTLINE [14] connects patient medical records with a medical literature search by identifying words in the chart that exist in any Metathesaurus term and selecting a search strategy. SAPHIRE [15] automates the indexing and retrieval of documents in medical literature databases to obtain better retrieval. Though these systems use different strategies and take different forms of input, all of them select a search strategy based on terms (or concepts).

Our approach extends this approach to include consideration of the relationships between the concepts. A basic hypothesis of our research is that many specific user queries can be approximated by a manageable number of "generic" queries [16]. Given an expressed information need, natural language processing is used to identify an appropriate generic query and to identify the appropriate concepts and relations to be used in the query. This enables us to generate search strategies in the context of an entire user query, instead of individual terms. If the number of generic queries is truly manageable, then it is feasible to develop, in advance, search strategies for each query which can be instantiated with appropriate user terms.

While the rules presently used by the search strategy generator cover all of the syntax of the BRS search engine, the semantic patterns take advantage of only certain features. Current work continues toward expanding the patterns needed for the variety of semantic triples encountered in generic queries. Further research is needed to determine how to apply other features such as term explosion and stemming.

There is ample opportunity to incorporate additional strategies into the knowledge-based approach. For example, the MeSH co-occurrence data from the UMLS could be used, much as it is in the Q&A component of IQW [17], to estimate the size of retrievals. Furthermore, post-processing rules could be added that can modify search strategies based on preliminary retrieval results. For example, if a strategy retrieves too many citations, additional search terms could be added to restrict the results; if too few citations are retrieved, a nonessential term could be dropped.

Although the current system works only with a single information source, the scheme of the computational model is independent of a particular information retrieval language or system. The syntax and related librarian knowledge are declarative and are separated from the computational procedure. With this approach, the system can generate information retrieval strategies for different systems, by supplying the corresponding syntax and indexing terms, without requiring significant change to the program.

CONCLUSION

This work explores a method for modeling expertise used by reference librarians when they assist users in information retrieval. The model is incorporated into a program that generates sophisticated search strategies based on information that was ultimately derived from the user's natural language. The use of conceptual graphs as an intermediate formalism separates the understanding of the question from the techniques needed to obtain an answer.

Acknowledgement

This work was supported by a Unified Medical Language System contract from the National Library of Medicine and by the IBM Corporation.

References

1. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The Medline Button. In Frisse ME, ed.: *Proceedings of the Sixteenth Annual SCAMC*; McGraw-Hill, New York, 1992:81-85.
2. National Library of Medicine. *UMLS Knowledge Sources - 3rd Experimental Edition.* Bethesda (MD): The Library, 1992.
3. Lindberg DAB and Humphreys BL. The UMLS knowledge sources: tools for building better user interfaces. In Miller RA, ed.: *Proceedings of the Fourteenth Annual SCAMC.* IEEE Computer Society Press, New York, 1990:121-125.
4. Tuttle MS, Sherertz D, Erlbaum M, Olson N and Nelson SJ. Implementing Meta-1: the first version of the UMLS Metathesaurus. In Kingsland LC, ed.: *Proceedings of the Thirteenth Annual SCAMC.* IEEE Computer Society Press, New York, 1989:483-487.
5. McCray AT and Hole WT. The Scope and Structure of the First Version of the UMLS Semantic Net. In Miller Ra, ed.: *Proceedings of the Fourteenth Annual SCAMC.* IEEE Computer Society Press, New York, 1990:126-130.
6. Johnson SB, Aguirre T, Peng P, and Cimino JJ. Interpreting Natural Language Queries Using the UMLS. In Safran C, ed.: *Proceedings of the Seventeenth Annual SCAMC*; McGraw-Hill, New York, 1993: (in press).
7. Sowa JF. *Conceptual Structures: Information Processing in Mind and Machine.* Reading (MA): Addison-Wesley, 1984.
8. Peirce CS. *Collected Papers of Charles Sanders Peirce.* Burks AW, editor.. Cambridge (MA): Harvard University Press, Vol.4:320-410.
9. Roberts DD. *The Existential Graphs of Charles S. Peirce.* The Hague: Mouton, 1973.
10. BRS Colleague User Manual. BRS Information Technologies. Inc. McLean, VA, 1992.
11. Sterling L and Shapiro E. *The Art of PROLOG.* Cambridge (MA): MIT Press, 1986.
12. Kingsland III LC, Harbourt AM, Syed EJ and Schuyler PL. Coach: applying UMLS Knowledge Sources in an expert searcher environment. *Bull Med Libr Assoc*; April 1993:81(2).
13. Cimino C, Barnett GO. Standardizing access to computer-based medical resources. In Miller RA (ed): *Proceedings of the Fourteenth Annual SCAMC*, IEEE Computer Society Press, New York, 1990:33-37.
14. Miller RA, Gieszczykiewicz FM, Vries JK, Cooper GF. CHARTLINE: providing bibliographic references to patient charts using the UMLS Metathesaurus knowledge sources. In Frisse ME, ed.: *Proceedings of the Sixteenth Annual SCAMC*; McGraw-Hill, New York, 1992:86-90.
15. Hersh W, Hickam DH, Haynes RB, McKibbon KA. Evaluation of SAPHIRE: An Automated Approach to Indexing and Retrieving Medical Literature. In Clayton PD, ed.: *Proceedings of the Fifteenth Annual SCAMC*; McGraw-Hill, New York, 1991:808-812.
16. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. *Bulletin of the Medical Library Association.* April 1993:81(2).
17. Merz RB, Cimino C, Barnett GO, Blewett DR, Gnassi JA, Grundmeier R, and Hassan L. Q & A: A Query Formulation Assistant. In Frisse ME, ed.: *Proceedings of the Sixteenth Annual SCAMC*; McGraw-Hill, New York, 1992:498-502.